

On the VC dimension of bounded margin classifiers

Don Hush and Clint Scovel*

Los Alamos National Laboratory,
Los Alamos, NM, 87545
(dhush@lanl.gov and jcs@lanl.gov)

April 13, 2000

Dedicated to Ané.

Abstract

In this paper we prove a result that is fundamental to the generalization properties of Vapnik's support vector machines and other large margin classifiers. In particular, we prove that the minimum margin over all dichotomies of $k \leq n + 1$ points inside a unit ball in R^n is maximized when the points form a regular simplex on a unit sphere. We also provide an alternative proof directly in the framework of level fat shattering.

* partially supported by The Los Alamos DOE Program in Applied Mathematical Sciences

1. Introduction

Bounds on the generalization error for classifiers that minimize empirical error are available as a function of the number of training samples and the complexity of the set of functions from which the classifier was drawn. One of the most widely used complexity measures is the Vapnik-Chervonenkis (VC) dimension, which for linear classifiers is $n+1$, where n is the dimension of the ambient space. VC generalization bounds support the conventional wisdom that the best way to control a classifier's complexity is to control its size. But this view does not account for the observation that the generalization of a fixed size classifier can often be improved by maximizing the amount by which it separates the data. One of the first such results is due to Vapnik who showed that the VC dimension of linear classifiers restricted to a particular data set can be bounded in terms of their *margin*, which measures how much they separate the data. This result is formalized by Theorem 1 below, whose proof is the main concern of this paper. Although Vapnik's theorem did not immediately yield generalization results (because of its data dependent nature, see Shawe-Taylor et. al. [10]) it suggested that the generalization of a large margin classifier could be controlled independent of its size (n), and provided a key motivation for Vapnik's Support Vector Machines [12]. More recently, Bartlett [1] and Shaw-Taylor, et. al. [10] have developed rigorous generalization bounds in terms of the margin by using a complexity measure called the fat-shattering dimension, which we discuss briefly in section 4. These bounds help explain the success of a number of recent approaches that are aimed at maximizing the margin, e.g. Support Vector Machines [12], Boosting [4], and Direct Optimization of Margin (DOOM) [6].

Definition 1 Let $X = \mathbb{R}^n$ be the n -dimensional Euclidean space, and let H be the family of linear classifiers $c(x) = \text{sign}(h(x))$ where $h(x)$ is an affine function. Further, let H_ρ be the set of linear classifiers that dichotomize X using hyperplanes of thickness ρ . More formally, define H_ρ to be classifiers of the form

$$c_\rho(x) = c(x), \quad D(x|h=0) > \frac{\rho}{2}$$

where $D(x|h=0)$ is the distance from x to the hyperplane $h=0$. (Note that $c_\rho(x)$ is not defined for $\{x : D(x|h=0) \leq \frac{\rho}{2}\}$.) The margin of classifiers in H_ρ is defined to be ρ . Finally, let $H_{\rho+}$ be the set of linear classifiers with thickness greater than or equal to ρ , that is $H_{\rho+} = \cup_{\phi \geq \rho} H_\phi$.

The SVM method produces classifiers of maximal margin that correctly classify a fixed size training set. The following theorem, due to Vapnik [11, 12], provides the essential link between margin and realized classifier class complexity for SVMs.

Theorem 1 (Vapnik, 1982) Let $X_r = \{x_1, x_2, \dots, x_k\} \subset X$ denote a set of points contained within a sphere of radius r . The VC dimension of $H_{\rho+}$ restricted to X_r satisfies

$$VCdim(H_{\rho+}) \leq \min(\lceil \frac{4r^2}{\rho^2} \rceil, n) + 1.$$

Actually, it is Burges [3] that noted that $\lceil \cdot \rceil$ = "smallest integer greater than or equal to" needs to be used in the statement of this theorem instead of the more often quoted $\lfloor \cdot \rfloor$ = "largest integer less than or equal to", which is incorrect, but asymptotically sharper. To prove this result it is sufficient to determine the largest set X_r that can be shattered by the set of hyperplanes of thickness ρ . The upper bound is obviously $n + 1$ (the number shattered when $\rho = 0$). Existing proofs of the (potentially) tighter bound, $\lceil (2r/\rho)^2 \rceil + 1$ rely on the almost obvious assumption that the minimum margin over all dichotomies of $k \leq n + 1$ points in \mathbb{R}^n can be maximized by placing these points on a regular simplex whose vertices lie on the surface of the sphere(See [11], page 324 or [12], page 353). Although this assumption has intuitive appeal, it has not been proven correct(cf. Burges [3]).

The purpose of this paper is to provide such a proof. Indeed we provide two proofs. The first is directly in terms of the margin and the second is in the framework of level fat shattering. Shawe-Taylor et. al. [10] observed the connection between level fat shattering and margin. Indeed, Shawe-Taylor et. al. [10] prove a bound in the *level fat shattering* formulation as a corollary to Vapnik's Theorem. On the other hand, Gurvits [5] provides a bound in the level fat shattering formulation which has a weaker bound than Vapnik's as a corollary. Bartlett and Shawe-Taylor [2] use Gurvits' idea to prove bounds on the level fat shattering dimension of homogeneous linear classifiers, but these bounds do not directly apply here because of the homogeneity assumption. For our second proof, we use a modification of the technique used by Gurvits [5] and Bartlett and Shawe-Taylor [2].

We begin by establishing the following lemma.

2. Preparation

Let $x = (x_1, x_2, \dots, x_k)$ denote a vector of k points in \mathbb{R}^n . Define $r(x)$ to be the radius of the smallest ball in \mathbb{R}^n that contains all k points.

Lemma 1 *Suppose that $r(x) \leq 1$. Then in the center of mass frame(translating the data so that $\sum_i x_i = 0$)*

$$\sum |x_i|^2 \leq k.$$

Proof. Let E denote averaging over the data index i . $r(x) \leq 1$ implies that $x_i = z + y_i$ where $|y_i| \leq 1$ for some z . Since the variance is translation invariant

$$E(|X - E(X)|^2) = E(|Y - E(Y)|^2) = E(|Y|^2) - E(Y)^2 \leq E(|Y|^2) \leq 1$$

and the proof is finished.

□

3. Statement and Proof of the Theorem

We now state and prove the main theorem.

Theorem 2 *Let $x = (x_1, x_2, \dots, x_k)$ denote a vector of k points in \mathbb{R}^n . Define $r(x)$ to be the radius of the smallest ball in \mathbb{R}^n that contains all k points. Let s denote a proper subset of the k integers $\{1, 2, \dots, k-1, k\}$ and let x_s denote the set of points corresponding to the subset s . Let $\rho(x_s)$ be the distance between the convex hull of x_s and the convex hull of its complement x_{s^c} .*

Then the value

$$\max_{x: r(x) \leq r} \min_s \rho(x_s)^2$$

is obtained when x is a regular simplex with vertices on the sphere of radius r .

Proof. We first note that since k points span at most a $k-1$ dimensional affine subspace, we can restrict to $n = k-1$. It is also clear that $\max_{x: r(x) \leq r} \min_s \rho(x_s)^2$ is quadratic in r , so we need to prove that the value

$$\max_{x: r(x) \leq 1} \min_s \rho(x_s)^2$$

is obtained when x is a regular simplex with vertices on the unit sphere. Define

$$h(k) \doteq \max_x \min_s \rho(x_s)^2$$

where x is constrained so that $r(x) \leq 1$ and s varies over all the proper subsets of the k points. This is a $\max_x \min_s$ game with payoff function $\rho(x_s)^2$, lower value $h(k) = \max_x \min_s \rho(x_s)^2$, and upper value $v(k) = \min_s \max_x \rho(x_s)^2 = 1$. In general, $h(k) \leq v(k)$.

Our plan of attack is as follows. We extend to a game with payoff function $f(x, y)$ with the same lower value. Then we explicitly construct a saddle point (x_0, y_0) to this extended game with x_0 a regular k -simplex, where a saddle point (x_0, y_0) satisfies $f(x, y_0) \leq f(x_0, y_0) \leq f(x_0, y)$ for all x and y . By von Neumann's Theorem (von Neumann and Morgenstern [8] pg 95.),

$$h(k) = f(x_0, y_0).$$

This proves the theorem.

To make x_s a vector we define $x_s = (x_{i_1}, x_{i_2}, \dots, x_{i_{|s|}})$, where i_j are all in s and they are monotonic $i_1 < i_2 < \dots < i_{|s|}$. Observe that $\rho(x_s)^2$ itself is a minimization

$$\rho(x_s)^2 = \min_{p^s, q^s} \left| \sum_{i \in s} p_i^s x_i - \sum_{j \in s^c} q_j^s x_j \right|^2$$

where p^s are vectors of length $|s|$, with $p_i^s \geq 0, i = 1, \dots, |s|$ and $\sum_{i=1, \dots, |s|} p_i^s = 1$ and likewise for q^s except that it is of length $|s^c| = k - |s|$. Therefore we first rewrite the max-min game as a max-min game with payoff function $F(x, z) =$

$|\sum_{i \in s} p_i^s x_i - \sum_{j \in s^c} q_j^s x_j|^2$ where $z = (s, p^s, q^s)$. We extend again by observing that

$$\min_{(s, p^s, q^s)} \left| \sum_{i \in s} p_i^s x_i - \sum_{j \in s^c} q_j^s x_j \right|^2 = \min_{(\lambda, p, q)} \sum_s \lambda_s \left| \sum_{i \in s} p_i^s x_i - \sum_{j \in s^c} q_j^s x_j \right|^2,$$

where λ varies over all probability distributions over the set of proper subsets and where $p = \prod_s \{p^s\}$ and $q = \prod_s \{q^s\}$ are the product variables. This forms a new min-max game with the same lower value as the original with payoff function $f(x, y) = \sum_s \lambda_s \left| \sum_{i \in s} p_i^s x_i - \sum_{j \in s^c} q_j^s x_j \right|^2$ where $y = (\lambda, p, q)$. Consequently the chain of extensions can be written

$$h(k) = \max_x \min_s \rho(x_s)^2 = \max_x \min_z F(x, z) = \max_x \min_y f(x, y)$$

Denote $\lfloor \cdot \rfloor = \lfloor \cdot \rfloor$ "the largest integer less then or equal to" operator.

Lemma 2 *The function*

$$f(x, y) = \sum_s \lambda_s \left| \sum_{i \in s} p_i^s x_i - \sum_{j \in s^c} q_j^s x_j \right|^2$$

has a saddle point at (t, y^*) where t is the regular simplex on the unit sphere and $y^* = (1_{\lfloor k/2 \rfloor}, P, Q)$ where $1_{\lfloor k/2 \rfloor}$ is the probability whose mass lies uniformly distributed over the set of subsets s such that $|s| = \lfloor \frac{k}{2} \rfloor$ and $P^s = \frac{1}{|s|}(1, 1, \dots, 1, 1)$ and $Q^s = \frac{1}{k-|s|}(1, 1, \dots, 1, 1)$.

Proof. Recall the definition of a saddle at (t, y^*) :

$$f(x, y^*) \leq f(t, y^*) \leq f(t, y)$$

for all x and y . We prove these inequalities one at a time.

Proof of $f(t, y^*) \leq f(t, y)$:

The simplex is special in that

$$\left| \sum_{i \in s} p_i^s t_i - \sum_{j \in s^c} q_j^s t_j \right|^2 = \sum_{i \in s} (p_i^s)^2 + \sum_{j \in s^c} (q_j^s)^2$$

which has its minimum value $\frac{k^2}{(k-1)|s|(k-|s|)}$ at $p^s = P^s$ and $q^s = Q^s$.

Consequently,

$$f(t, (\lambda, P, Q)) \leq f(t, (\lambda, p, q)).$$

Since the function $\frac{k^2}{(k-1)|s|(k-|s|)}$ is constant on the strata of subsets of size $|s|$,

$$f(t, (\lambda, P, Q)) = \frac{k^2}{k-1} \sum_s \frac{1}{|s|(k-|s|)} \lambda_s = \frac{k^2}{k-1} \sum_{|s|} \lambda_{|s|} \frac{1}{|s|(k-|s|)}.$$

Since $\frac{1}{|s|(k-|s|)}$ is minimal at $|s| = \lfloor \frac{k}{2} \rfloor$, $f(t, (\lambda, P, Q))$ is then minimized by placing the mass of λ entirely on $|s| = \lfloor \frac{k}{2} \rfloor$. Consequently,

$$f(t, (1_{\lfloor k/2 \rfloor}, P, Q)) \leq f(t, (\lambda, P, Q)),$$

and therefore

$$f(t, y^*) \leq f(t, y).$$

Proof of $f(x, y^*) \leq f(t, y^*)$:

By definition

$$f(x, y^*) = \frac{1}{\binom{k}{\lfloor k/2 \rfloor}} \sum_{s: |s| = \lfloor \frac{k}{2} \rfloor} \left| \frac{1}{\lfloor \frac{k}{2} \rfloor} \sum_{i \in s} x_i - \frac{1}{k - \lfloor \frac{k}{2} \rfloor} \sum_{i \in s^c} x_i \right|^2,$$

and since the constraint is translation invariant we translate to the center of mass so that $0 = \sum x_i$. Consequently $f(x, y^*)$ is a positive multiple of

$$\sum_{s: |s| = \lfloor \frac{k}{2} \rfloor} \left| \sum_{i \in s} x_i \right|^2.$$

Reverse the order of summation and expand so that

$$\sum_{s: |s| = \lfloor \frac{k}{2} \rfloor} \left| \sum_{i \in s} x_i \right|^2 = \sum_{i, j} \sum_{s: |s| = \lfloor \frac{k}{2} \rfloor, \{i, j\} \subset s} x_i \cdot x_j.$$

The interior sum $\sum_{s: |s| = \lfloor \frac{k}{2} \rfloor} x_i \cdot x_j$ is $x_i \cdot x_j$ times the number of subsets which contain both i and j . When $i = j$ it is $\binom{k-1}{\lfloor \frac{k}{2} \rfloor - 1}$ but when $i \neq j$, $\binom{k-2}{\lfloor \frac{k}{2} \rfloor - 2}$. Consequently,

$$\begin{aligned} \sum_{i, j} \sum_{s: |s| = \lfloor \frac{k}{2} \rfloor, \{i, j\} \subset s} x_i \cdot x_j &= \binom{k-1}{\lfloor \frac{k}{2} \rfloor - 1} \sum_i |x_i|^2 + \binom{k-2}{\lfloor \frac{k}{2} \rfloor - 2} \sum_{i \neq j} x_i \cdot x_j \\ &= \left(\binom{k-1}{\lfloor \frac{k}{2} \rfloor - 1} - \binom{k-2}{\lfloor \frac{k}{2} \rfloor - 2} \right) \sum_i |x_i|^2 + \binom{k-2}{\lfloor \frac{k}{2} \rfloor - 2} \sum_{i, j} x_i \cdot x_j, \end{aligned}$$

but since $0 = \sum x_i$ and $\binom{k-1}{\lfloor \frac{k}{2} \rfloor - 1} > \binom{k-2}{\lfloor \frac{k}{2} \rfloor - 2}$ the second term vanishes and we are left with a positive multiple of

$$\sum_i |x_i|^2$$

From Lemma 1, we know that

$$\sum_i |x_i|^2 \leq k$$

and for the simplex t

$$\sum_i |t_i|^2 = k.$$

Therefore,

$$f(x, y^*) \leq f(t, y^*).$$

The proof of Lemma 2 and therefore of Theorem 2 is finished.

□

4. Level fat shattering

As mentioned in the introduction, Theorem 1 assumed the bounds between the number of data points and the margin for the regular simplex provided such bounds in general. Explicit calculation on a regular unit simplex gives

$$\rho \leq \frac{2}{\sqrt{k-1}} \quad k \text{ even}$$

$$\rho \leq \frac{2k}{k-1} \frac{1}{\sqrt{k+1}} \quad k \text{ odd},$$

and the bounds stated in Theorem 1 represent implied inversions to bounds of k in terms of ρ . Shawe-Taylor et.al. [10] discuss why Vapnik's theorem does not provide bounds on generalization error, even though it did provide motivation for Support Vector Machines and other large margin classifiers. They resolve this issue by utilizing the level fat shattering formalism. In particular they show that if the data is γ level fat shattered then each partition can be achieved by a hyperplane with margin $\rho \geq 2\gamma$. Consequently, Theorem 2 applies with $\rho = 2\gamma$. Bartlett and Shawe-Taylor [2] use Gurvits' idea to prove bounds on the level fat shattering dimension of homogeneous linear classifiers, but these bounds do not directly apply here because of the homogeneity assumption. We now define fat shattering and level fat shattering and prove the equivalent to Theorem 2 using the level fat shattering formalism directly.

Definition 2 *k points are γ fat shattered by the affine linear functions if there is a $\alpha_i, i = 1, \dots, k$ such that for each partition b , there is a choice of unit vector ω_b and a ϕ_b so that*

$$\omega_b \cdot x_i + \phi_b \geq \alpha_i + \gamma, \quad b_i = 1$$

$$\omega_b \cdot x_i + \phi_b \leq \alpha_i - \gamma, \quad b_i = -1.$$

In level fat shattering, we require $\alpha_i = \alpha$ to be constant. Then the constant α can be absorbed in ϕ_b so we can set it to zero in the formulation as follows:

$$\omega_b \cdot x_i + \phi_b \geq \gamma, \quad b_i = 1$$

$$\omega_b \cdot x_i + \phi_b \leq -\gamma, \quad b_i = -1.$$

which can be written in the concise form

$$b_i(\omega_b \cdot x_i + \phi_b) \geq \gamma.$$

Theorem 3 *Suppose that k points are contained in a ball of radius 1 ($r(x) \leq 1$) and are γ level fat shattered by the affine linear functions. Then*

$$\gamma \leq \frac{1}{\sqrt{k-1}} \quad k \text{ even}$$

$$\gamma \leq \frac{k}{k-1} \frac{1}{\sqrt{k+1}} \quad k \text{ odd}.$$

Because these bounds are those of a regular simplex, application of the result of Shawe-Taylor [10] sending $2\gamma \rightarrow \rho$ shows that Theorem 3 provides an alternative proof of Theorem 2.

Proof. First consider k even. For a subset S_b of size $|S_b| = \frac{|S|}{2}$, we have $\sum_i b_i = 0$, so it follows that

$$\omega_b \cdot \sum_i b_i x_i \geq \gamma |S|,$$

which by the Cauchy-Schwartz inequality implies

$$|\sum_i b_i x_i| \geq \gamma |S|.$$

In Bartlett and Shawe-Taylor [2] the next step is to average $|\sum_i b_i x_i|^2$ over all b but we instead average over all b with $|S_b| = \frac{|S|}{2}$. By reversing the order of summation

$$E(|\sum_i b_i x_i|^2) = \sum_{i,j} E(b_i b_j) x_i \cdot x_j,$$

we apply standard combinatorial arguments to determine the coefficients of $x_i \cdot x_j$. Since $b_i b_j = 1$ when $i = j$, the constant in front of $\sum_i x_i^2$ is one. For $i \neq j$

$$E(b_i b_j) = 2 \frac{\binom{\frac{k-2}{2}-2} - \binom{\frac{k-2}{2}-1}}{\binom{\frac{k}{2}}{2}} = -\frac{1}{k-1}.$$

Then

$$\begin{aligned} E(|\sum_i b_i x_i|^2) &= \sum_i x_i^2 - \frac{1}{k-1} \sum_{i \neq j} x_i \cdot x_j \\ &= \frac{k}{k-1} \sum_i x_i^2 - \frac{1}{k-1} \sum_{i,j} x_i \cdot x_j = \frac{k}{k-1} \sum_i x_i^2 - \frac{1}{k-1} |\sum_i x_i|^2 \\ &\leq \frac{k}{k-1} \sum_i x_i^2. \end{aligned}$$

Consequently, using the center of mass frame, Lemma 1 implies that

$$E(|\sum_i b_i x_i|^2) \leq \frac{k^2}{k-1}$$

so that for at least one subset with $|S_b| = \frac{|S|}{2}$,

$$|\sum_i b_i x_i|^2 \leq \frac{k^2}{k-1}.$$

Combining this with the previous bound $|\sum_i b_i x_i| \geq \gamma|S|$ gives

$$\gamma \leq \frac{1}{\sqrt{k-1}}$$

completing the proof when k is even.

Now consider k odd. Let b denote a partition such that $|S_b| = \lfloor \frac{|S|}{2} \rfloor + 1$. Denote $r = |S_b|$. Then $r = \lfloor \frac{k}{2} \rfloor + 1$ and $k = 2r - 1$. We wish to proceed as in the even case above but must modify the procedure so that summing the inequality

$$b_i(\omega_b \cdot x_i + \phi_b) \geq \gamma$$

cancels the unknown ϕ_b . We accomplish this by defining the weights

$$L(b)_i = \frac{1}{r} \quad b_i = 1$$

$$L(b)_i = \frac{1}{r-1} \quad b_i = -1.$$

Then $\sum_i b_i L(b)_i = 0$ and $\sum_i L(b)_i = 2$. Since L is positive we multiply the shattering equations by $L(b)_i$ and sum to obtain

$$\omega_b \cdot \sum_i b_i L(b)_i x_i \geq 2\gamma,$$

which by the Cauchy-Schwartz inequality implies that

$$|\sum_i b_i L(b)_i x_i| \geq 2\gamma.$$

We now compute an upper bound of

$$E(|\sum_i b_i L(b)_i x_i|^2) = \sum_{i,j} E(b_i b_j L(b)_i L(b)_j) x_i \cdot x_j$$

where E denotes averaging over all partitions b such that $|S_b| = \lfloor \frac{|S|}{2} \rfloor + 1$. We first consider the case when $i = j$. When $i \in S_b$, $L(b)_i = \frac{1}{r}$ and $L(b)_i = \frac{1}{r-1}$ otherwise. Consequently,

$$E(b_i b_j L(b)_i L(b)_j) = \frac{\frac{1}{r^2} \binom{2r-2}{r-1} + \frac{1}{(r-1)^2} \binom{2r-2}{r}}{\binom{2r-1}{r}}$$

and a little computation yields

$$E(b_i b_j L(b)_i L(b)_j) = \frac{1}{r(r-1)}.$$

Now consider when $i \neq j$. Then

$$b_i b_j L(b)_i L(b)_j = \begin{cases} \frac{1}{r^2}, & \{i, j\} \subseteq S_b \\ \frac{1}{(r-1)^2}, & \{i, j\} \subseteq S_b^c \\ -\frac{1}{r(r-1)}, & (i \in S_b, j \in S_b^c) \text{ or } (i \in S_b^c, j \in S_b). \end{cases}$$

Consequently

$$E(b_i b_j L(b)_i L(b)_j) = \frac{\frac{1}{r^2} \binom{2r-3}{r-2} + \frac{1}{(r-1)^2} \binom{2r-3}{r} - \frac{2}{r(r-1)} \binom{2r-3}{r-1}}{\binom{2r-1}{r}}$$

and a little computation yields

$$E(b_i b_j L(b)_i L(b)_j) = -\frac{1}{2r(r-1)^2}.$$

Consequently,

$$\begin{aligned} E(|\sum_i b_i L(b)_i x_i|^2) &= \frac{1}{r(r-1)} \sum_i x_i^2 - \frac{1}{2r(r-1)^2} \sum_{i \neq j} x_i \cdot x_j \\ &= \frac{2r-1}{2r(r-1)^2} \sum_i x_i^2 - \frac{1}{2r(r-1)^2} \sum_{i,j} x_i \cdot x_j \\ &= \frac{2r-1}{2r(r-1)^2} \sum_i x_i^2 - \frac{1}{2r(r-1)^2} |\sum_i x_i|^2 \end{aligned}$$

which is bounded by

$$\frac{|2r-1|^2}{2r(r-1)^2}$$

by moving to the center of mass frame, applying Lemma 1, and recalling that $k = 2r - 1$.

Consequently there must exist such a partition b so that

$$|\sum_i b_i L(b)_i x_i|^2 \leq \frac{|2r-1|^2}{2r(r-1)^2}$$

but since

$$|\sum_i b_i L(b)_i x_i| \geq 2\gamma$$

we obtain

$$\gamma^2 \leq \frac{|2r-1|^2}{8r(r-1)^2} = \frac{k^2}{(k-1)^2(k+1)}$$

or

$$\gamma \leq \frac{k}{k-1} \frac{1}{\sqrt{k+1}},$$

and the proof is finished. □

5. Epilogue

Shawe-Taylor et.al. [10] have shown that a map from maximizing the margin over all labelings of the data to finding the greatest γ so that the data is γ level fat shattered is achieved by $\rho \rightarrow 2\gamma$. We have proven Vapnik's theorem both in the original margin formulation and in the equivalent level fat shattering formulation. The relationship between these proofs is unclear. In the first proof there was no need to know the value of the margin for the regular simplex while in the second there was. We note that the choice of $L(b)$ used in the second proof can also be seen in the first technique. We suspect that this choice can be justified by formulating the maximum level fat shattering problem as a convex programming problem, extending this problem to a Lagrangian which has a saddle point at the solution and then utilizing the saddle point property of the solution in much the same way as we did in the first proof technique. It would be useful to better understand how the transformation of optimization problems from margin to γ level fat shattering induces transformations of proof techniques. We also suspect that the regular simplex is the only configuration that achieves equality in the inequalities of Theorem 3.

Acknowledgments

We would like to express our thanks to the referees for many helpful suggestions, including the encouragement to complete the proof for the level fat shattering framework. Thanks also to the referee who informed us of an elementary proof of Lemma 1, which we have subsequently adopted.

References

1. Bartlett, P. L., The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network, *IEEE Transactions on Information Theory* **44**(1998), 525–536.
2. Bartlett, P.L., Shawe-Taylor, J., Generalization performance of support vector machines and other pattern classifiers, *Advances in Kernel Methods: Support Vector Learning*, pp 43–54, Schölkopf, B., Burges, C. J. C., and Smola, A. J., Eds., MIT Press, 1999.
3. Burges, C. J. C., A tutorial on Support Vector Machines for pattern recognition, *Data Mining and Knowledge Discovery* **2**(1998), 121–167.
4. Freund, Y., and Schapire, R. E., A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory: Second European Conference, EuroCOLT '95*, pp. 23–37., Springer-Verlag, 1995. See also *Journal of Computer and System Science*, **55**(1) (1997), 119–139.
5. Gurvits, L., A note on the scale sensitive dimension of linear bounded functionals in Banach spaces, *Proceedings of Algorithm Learning Theory*, **ALT-97**(1997).
6. Mason, L., Bartlett, P.L., and Baxter, J., Direct optimization of margins improves generalization in combined classifiers, to appear in *Advances in Neural Information Processing Systems* **12**, MIT Press, Cambridge, MA (1999).
7. von Neumann, J., Zur Theorie der Gesellschaftspiele, *Mathematische Annalen* **100**(1928), 295–320.
8. von Neumann, J., and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, 1944.

9. Polak, E., *Optimization: Algorithms and Consistent Approximations*, Springer-Verlag, New York, 1997.
10. Shawe-Taylor, J., Bartlett, P.L., Williamson, R. C., and M. Anthony, Structural Risk Minimization over Data-Dependent Hierarchies, *Neuro COLT Technical Report NC-TR-96-053*(1996).
11. Vapnik, V., *Estimation of Dependencies Based on Empirical Data*, translated by S. Kotz, Springer-Verlag, New York, 1982.
12. Vapnik, V. N., *Statistical Learning Theory*, John Wiley and Sons, Inc., New York, 1998.